Modelling and Influencing the AI Bidding War: A Research Agenda

The Anh Han, Luis Moniz Pereira, Tom Lenaerts

Abstract

A race for technological supremacy in AI could lead to serious negative consequences, especially whenever ethical and safety procedures are underestimated or even ignored, leading potentially to the rejection of AI in general. For all to enjoy the benefits provided by safe, ethical and trustworthy AI systems, it is crucial to incentivise participants with appropriate strategies that ensure mutually beneficial normative behaviour and safety-compliance from all parties involved. Little attention has been given to understanding the dynamics and emergent behaviours arising from this AI bidding war, and moreover, how to influence it to achieve certain desirable outcomes (e.g. AI for public good and participant compliance). To bridge this gap, this paper proposes a research agenda to develop theoretical models that capture key factors of the AI race, revealing which strategic behaviours may emerge and hypothetical scenarios therein. Strategies from incentive and agreement modelling are directly applicable to systematically analyse how different types of incentives (namely, positive vs. negative, peer vs. institutional, and their combinations) influence safety-compliant behaviours over time, and how such behaviours should be configured to ensure desired global outcomes, studying at the same time how these mechanisms influence AI development. This agenda will provide actionable policies, showing how they need to be employed and deployed in order to achieve compliance and thereby avoid disasters as well as loosing confidence and trust in AI in general.

Introduction

Research and development in different areas of fundamental and applied Artificial Intelligence (AI) have been making encouraging progress. Within the research community, there is a growing effort to make progress towards Artificial General Intelligence (AGI). AI brings enormous potential benefits across many sectors, being recognised as a strategic priority by a range of actors and stakeholders, including representatives of various businesses, private research groups, companies, and governments (AI-Roadmap-Institute 2017b). The media attention as well as the (un)announced business and political ambitions indicate that an AI bidding war has been triggered, competing to be the first to develop and deploy a powerful, transformative AI (Armstrong, Bostrom, and Shulman 2016; Baum 2017; Bostrom 2017; Cave and Ó hÉigeartaigh 2018). These AI systems could be either AGI, able to perform a broad set of intellectual tasks while continually improving itself, sufficiently powerful specialised AIs or even AI specifically developed for espionage and cyberterrorism.

An AI race for technological advantage towards powerful AI systems could lead to serious negative consequences, especially when ethical and safety procedures are underestimated or even ignored (Armstrong, Bostrom, and Shulman 2016; Cave and Ó hÉigeartaigh 2018). Safety and ethical agreements and regulations can be adopted to ensure that all parties involved in the race will comply with a set of mutually agreed standards and norms (Shulman and Armstrong 2009; Shulman and Armstrong 2009). However, as experience with many international treaties, such as climate change, fisheries and timber agreements (Barrett 2016; Cherry and McEvoy 2013; Nesse 2001) has shown, the autonomy and sovereignty of the parties involved will make monitoring and compliance enforcement difficult (if not impossible). Therefore, for all to enjoy the benefits provided by a safe, ethical and trustworthy AI, it is crucial to enact appropriate incentivising strategies in order to ensure mutual benefits and safety-compliance from all sides involved.

This position paper sets out a research agenda (i) to develop theoretical models (both analytic and simulated) that capture key factors of an AI race, in order to provide understanding on the dynamics and emergent strategic behaviours for different hypothetical race scenarios; and furthermore, (ii) to examine how incentives can be used to ensure desired outcomes and equilibria in this race. To this end, we combine research on incentives and agreement modelling, with the dynamical approaches that analyse population-wide dynamics, typically employed in Evolutionary Game Theory (EGT) (Sigmund 2010; Hofbauer and Sigmund 1998). Together, they will permit us to systematically explore how different types of incentives (namely, positive vs. negative, peer vs. institutional, and their combinations) can influence safety-compliance behaviours over time, and how such behaviours should be configured to ensure desired global outcomes (i.e. high levels of safety-compliance, possibly including compliant information sharing).

This research agenda is expected to lead to specific outcomes concerning the AI bidding war we are currently ob-

serving. On the one hand, it will provide methodologies for the investigation of AI race dynamics, including intervention mechanisms to influence outcomes. On the other, it will provide systematic understanding on how to promote safety compliance and the production of AI systems that benefit the many as opposed to the few. Moreover, conclusions derived from this project have the potential to resolve issues in other multiparty interaction domains, such as business transactions and international environmental agreements, for which commitments have been proven crucial (Singh 2013; Nesse 2001).

The rest of the paper is structured as follows. The next section describes relevant factors influencing the AI race and existing approaches for agreements and incentives modelling. It then describes our research agenda, which includes three directions aiming to provide a methodolgy for AI race modelling. Some preliminary results will be briefly described, elaborated further in the Supplementary Information section.

Background and Challenges

Below we review relevant issues and potential factors in AI race modelling, and EGT (Sigmund 2010) research on incentives, commitment and agreement modelling.

AI race and modelling

Potential AI disaster scenarios are many (Armstrong, Bostrom, and Shulman 2016; Pamlin and Armstrong 2015). However the uncertainties in accurately predicting these outcomes are high. In general, the first AGI are likely to be extremely powerful and, if not developed and controlled properly, might not be guaranteed human friendly (Armstrong, Bostrom, and Shulman 2016; Bostrum 2014). Those that first successfully develop more powerful AI will have significant benefits over others, and even 'the winner takes all' scenario might come about (AI-Roadmap-Institute 2017b). It is generally agreed that an AI-related disaster is more likely to occur when safety measures are ignored more often, which might be fostered by the speed and competition of the race. The risk of AI-related disaster increases when teams/developers do not devote sufficient attention and resources to safety in such a powerful system (AI-Roadmap-Institute 2017b) as a result of the on-going race's pressure.

Regardless of this disaster risk, people and teams can have different (biased) perceptions or beliefs about the level and nature of the risk as well as adopt different risk-taking behaviours (e.g. with different levels of risk-aversion). A high belief in risk might lead to fear mongering concerning both AI and AGI, thereby leading to overregulation and unnecessary obstruction to the development of AI and the realisation of social/economic benefits. On the other hand, a low belief in risk might lead to more risk-taking behaviours, and hence safety measures being omitted more often. Furthermore, envisaging the AI race at diverse temporal scales is likely to bring about distinct aspects that should be focused on (AI-Roadmap-Institute 2017a). For instance, each team might anticipate different speeds of reaching the first general AI system. A low belief about fast AGI arrival could result in miscalculating the risks of unsafe AGI deployment by non-compliant rogue teams.

Other important strategic aspects in the AI development race concern different types of openness, including openness about source code, science, data, safety techniques, capabilities, and goals (Bostrom 2017). Some forms of openness, e.g. about safety measures, are plausibly positive both short-term and long-term. Others, such as openness about source code, science, and possible capabilities, could intensify the competitive situation. A central concern is that openness could exacerbate a racing dynamic: competitors trying to be the first to develop true AIs may accept higher levels of (existential) risk in order to accelerate their progress. On the other hand, openness may reduce the probability of AI benefits being monopolized by a small group and facilitate counter measures.

Given the several factors and biases driving the strategic safety-compliant behaviours and the dynamics of interactions among development teams, the outcome of the race (e.g. whether or not disaster occurs) is difficult to predict. Very few efforts have been made in modelling AI disaster and race dynamics among teams. In (Armstrong, Bostrom, and Shulman 2016), the authors provide a game-theoretical model of an AI race, under the assumption that the first AI will be very powerful and transformative, so that each team is encouraged to finish first and thus skimp on safety precautions. This work, however, does not consider dynamical aspects of the race (e.g. how teams might adapt their strategies over time), and which strategic behaviours emerge in different race scenarios. Also, this work does not study how positive or negative incentives can be used to enhance teams's safety compliance.

To the contrary, there has been a significant body of computational modelling research regarding game-theoretical and EGT analyses of other disasters, such as climate change and nuclear war (Santos and Pacheco 2011; Baliga and Sjöström 2004). However, the AI race and its related risks are quite unique, according to an analysis of 12 large global catastrophic risks in (Pamlin and Armstrong 2015). Climate change disaster analyses primarily focus on the unwillingness of participants to take on themselves a personal cost for a jointly desired collective target, and is conjoined to collective risk by all parties involved (Santos and Pacheco 2011). In contrast, in AI racing, the winner(s) will derive significant relative advantage over others, risk being more individualized. The AI race is also different from the nuclear arms race, in that the former potentially poses greater effective achievement or otherwise risks to its creator, whereas nuclear powers are generally not at direct risk originating from their own arsenals (Armstrong, Bostrom, and Shulman 2016).

Finally, despite a number of proposals and debates on how to prevent, regulate, or solve the AI race, there is significant lack of rigorous modelling studies (Baum 2017; Cave and Ó hÉigeartaigh 2018; Geist 2016; Shulman and Armstrong 2009; Taddeo and Floridi 2018). Our proposed research agenda aims to bridge this gap by providing basic models to systematically understand the dynamics of the race and what strategic behaviours would likely emerge, in

different hypothetical scenarios and conditions of the race concerning disaster risk, risk-perception behaviours, level of information openness, number of racing teams, incentives deployment, etc. These models can then be used and perfected in discussions with stakeholders as well as policy makers.

Reward and punishment

Punishment and reward are major forms of incentives, widely adopted for enforcing cooperative behaviours among self-interested agents and enacting norm compliance in both social interactions and computerised systems (Sigmund, Hauert, and Nowak 2001; Herrmann, Thöni, and Gächter 2008; Chen et al. 2015). Various forms of such incentives have been studied, which can be roughly categorised into peer and institutional ones.

To provide a peer incentive, agents pay a personal cost to punish a violator (peer punishment) or reward a cooperator (peer reward), after an interaction. As a result, the punished violator and rewarded cooperator incur a decrease or increase, respectively, in their payoff. In the context of AI race behaviours, examples of peer incentives use are when teams refuse to support, and so do not share development progress and knowledge with non-compliant teams, consequently slowing down their development. They can also arrange, for instance, cyber-attacks against non-compliant teams, or even ascribe and spread their bad reputation, leading to those teams being unable to recruit and retain developers. On the other hand, highly compliant teams can be rewarded, obtaining more support, such as knowledge and experience sharing, and thus move faster along in the development race.

In contrast to peer incentives, institutional incentives assume instead the existence of an institution (with a budget) to take care of the incentivising process (Chen et al. 2015). Pool and coordinated group incentives, where agents can put together a fund before an interaction occurs to provide incentives subsequent to the interaction, can be considered as a first step towards institutionalisation of incentives (Sigmund et al. 2010; Boyd, Gintis, and Bowles 2010). Examples of institutions providing incentives are modern courts and policing systems, as well as international organisations such as the United Nations (UN), the European Union (EU), the World Trade Organization (WTO) or the Organisation for Economic Co-operation and Development (OECD). Setting-up and maintaining institutions are costly, but the presence of a powerful authority can effectively restrict individuals' strategic options and provide stronger incentives (Hilbe et al. 2014). In the context of an AI race, a centralised access to AI-related knowledge, algorithms and tools, as in the EU platform call (H2020-ICT-2018-2020) for instance, might provide institutional incentives such as strong support and punishment (e.g. allowing high levels of access or exclusion from the centralised pool of AI knowledge).

For both peer and institutional incentives, the critical conditions for cooperation to be achieved and sustained in evolutionary models, as well as observed in lab experiments, require that incentives 'fit the crime' (Sigmund et al. 2010; Boyd, Gintis, and Bowles 2010). That is, they increase with the severity of the violation or the merit of the cooperation. Moreover, they need to be cost-effective, that is, the effect of an incentive for its receiver should be sufficiently large compared to the cost to the provider. Interestingly, for institutional incentives, the centralisation by which an institution might observe a global population state enables more efficient approaches. For instance, the 'first carrot, then stick' policy that switches the incentive from reward to punishment whenever the frequency of cooperators in a population exceeds a threshold, is highly efficient at establishing full cooperation, better than either reward or punishment alone [21]. Furthermore, local institutions that provide incentives based on local neighbourhood properties (such as its cooperativeness level) have proved more efficient than large global governance overseeing the whole population, in the context of climate change games and cooperation dilemmas (Vasconcelos, Santos, and Pacheco 2013; Han et al. 2018). It may be that recourse to costly institutional supervision is only enacted when spontaneous peer cooperation becomes insufficient or inadequate.

Commitment modelling in dynamical systems

Commitment-based formalisms have been widely adopted in Multi Agent Systems (MAS) as modelling and engineering tools (Singh 2013), with important applications in business protocols, transactions, and software-oriented architectures. Representing commitments that agents have to one another specifies their expected interaction and correct behaviour (e.g. cooperation and norm compliance). The formalisms allow flexible capturing of contractual relationships amongst the agents concerned and, as such, incentivise their correct desirable behaviour. In a commitment-based MAS, it is crucial to understand how an adopted commitment formalism, including its compliance incentivising approach, influences agents' behaviours and the dynamics of their interactions and, as a result, determines whether it can actually help foster and secure substantial commitment compliance and correct behaviour. This is typically studied in the literature on dynamical MAS, using methods from EGT. In this dynamical context, agents interact and adapt their behaviours over time, e.g. via social learning (Sigmund 2010; Sigmund et al. 2010; Han, Pereira, and Lenaerts 2016).

In fact, commitments have been shown to provide important pathways to reaching mutual cooperation in the context of social dilemma interactions (namely, in the Prisoners Dilemma (PD) (Han et al. 2013) and Public Goods Game (PGG) (Han, Pereira, and Lenaerts 2016) settings), within populations of self-interested agents. This dynamical approach has provided important insights into the design of commitment-based MAS, enabling identification of areas of strengths and weaknesses of commitment-based mechanisms. Consequentially, it is now possible to make specific efforts to improve on concrete issues. This approach also provides novel understanding towards the long quest for the evolution of commitments and their roles in the evolution of cooperation (Nesse 2001); with important practical ramifications for social and economic interactions, ranging from personal relationships to business contracts and to international agreements (Barrett 2016; Cherry and McEvoy 2013).

However, these modelling studies focus on standard cooperation dilemmas settings, i.e. the PD and PGG (for both one-shot and repeated interactions). When moving to AI race interactions, several factors need to be taken into account, as described above, such as the risk of disaster, risk perception and levels of openness. And the incentives need to be accordingly designed to account for these new factors. For instance, by requiring that signatories to an AI race treaty or agreement will commit to allow internationally governed inspections concerning a defined minimal transparency or the detection of secret of undertakings. Our study may incorporate specific distinct punishments or rewards for the case of violation or otherwise, or lack of due diligence, in both such cases. Public and employee opinion and consumer practice can also be mustered as another form of incentive.

Research Agenda: problems to be studied

The literature reviewed above clearly shows the importance and gaps in modelling research to understand the dynamics of interactions and emergent behaviours in an AI race, as well as how incentive mechanisms might be used effectively to promote desired behaviours, such as safety standards and norms compliance from all parties involved in the race.

Our research agenda below aims to bridge this gap, and the following three directions or parts are envisaged to achieve that goal. The first one will develop baseline models for an AI race, on the basis of which the incentives for influencing the race will be studied in subsequent projects.

Development of baseline AI race models

The first direction aims to develop new, baseline strategic decision-making models that capture key factors relevant to the competition and cooperation among AI teams, and that investigate how they influence the outcome of the AI race. Namely, we will develop EGT models to answer the following questions:

- Which strategic behaviours emerge in the AI race?
- How does increasing the number of competing teams influence the evolutionary dynamics and outcome of the race, namely in preventing disaster?
- How does the risk-perception probability of AI disaster influence the race outcome?
- How does the heterogeneity of teams' development capacities influence the race outcome and strategic behaviours?

We envisage multi-player game models of multiple stages and/or repeated interactions, where a team might react to the development of strategic behaviours of other teams. The first teams to successfully develop AI will have significant benefits over others, and even the 'winner takes all' scenario might come about. Although we focus on multi-player games, for there are most likely multiple competing AI teams, as the race evolves the game might end up at a later stage with only a few or even two (strongest) teams. Our analysis will start with pair-wise interactions, modelled as two-player games, which will afterwards be generalised to multi-team interactions, modelled as multiplayer games, in order to provide a more comprehensive understanding.

We start with the most basic scenario where at each round of the race a team or player is faced with two possible choices: to follow the safety precaution (SAFE) or to ignore this safety precaution (UNSAFE). Since it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier, but also implies a slower development speed, compared to playing UNSAFE. As a generalisation of this binary-choice model we will consider continuous games where a player can choose the level of safety-precaution to adopt (i.e. SAFE and UNSAFE correspond to the two extreme cases of complete precaution and none at all, respectively).

Teams can collect benefits from their intermediate AI products. However, differently from the standard repeated games (Sigmund 2010) where all players collect benefit at every round, we will need to consider a new time scale, where different teams might collect benefits at different speeds. That would mean a possible time delay in players' decision-making, during the course of a repeated interaction, because they might want to wait for the outcome of a co-player's decision to see what choice he/she has adopted and/or will adopt in the next development round. Thus, a player has to decide whether to make an immediate move based on just present information (and hence be quicker to collect the next benefit and move faster in the race) but at the risk of making a worse choice, different from one that would have been chosen had the player already known the co-player's decision. Moreover, since noise is a key factor driving the emergent strategic behaviours in the context of repeated games (Sigmund 2010)—for instance when a team might (non-deliberately) make a mistake in the safety process, which might intensify the on-going race and trigger long-term retaliation (Martinez-Vaquero et al. 2015)—we will consider conflict resolution mechanisms such as apology and forgiveness (Martinez-Vaquero et al. 2015; McCullough 2008) for simmering down the noise effects on the race.

There will be a perception probability (by teams) that an AI disaster will occur wherein all teams lose the race, incurring a significant reduction in their payoff, and this risk-probability will follow a certain probability distribution. It is natural to assume that this probability increases with the frequency with which teams violate the safety requirements (i.e. play UNSAFE). Given the disaster probability, teams might have a different perception of the risk. For example, risk-taking (risk-avoiding) teams might underestimate (overestimate) this risk, leading to more (less) violations of the safety requirements. As in the case of climate change games (Santos and Pacheco 2011), the perception of risk is a key factor driving the evolutionary outcome (whether or not disaster happens). The main difference is that in climate change models the risk (of failure in avoiding disaster) is collective, whereas in the AI race risk is more individualised: should an individual team ignore safety requirements too often, the more likely their AI product will lead to disaster.

Additionally, assuming that teams have distinct development capacities, i.e. that they might move at different speeds

in the race, how does that change the strategic behaviours in the race by the different teams? Stronger teams might want to spend more effort with safety to guarantee no disaster occurs and ensure all the benefits from a powerful AI. On the other hand, weaker teams might want to water down or cease the safety efforts in order to catch up with the stronger teams.

Peer and group incentives for safety-compliance

This part of the agenda strives to investigate how peer incentives, such as peer punishment and reward, can be efficiently used, whether separately or jointly, for enhancing safety-compliance behaviours in the AI race games developed in the first part. Namely, we will address the following questions:

- What is the influence of using different types of peer incentives (separately or jointly), on safety-agreement compliance in the two-team AI race game?
- Generalising to multi-team agreements, how can peer incentives be used efficiently, taking into account the level of participation in the agreements as well as in a coordinated manner (i.e. pool and coordinated incentives)?
- How should incentives be customised to account for the level of risk (that an AI disaster will occur) as well as for teams with different capabilities of development and of commitment?

We shall start by extending the regimented agreement models for pair-wise and multiparty games, in both one-shot and repeated games, introduced in previous works (Han et al. 2013; Han, Pereira, and Lenaerts 2016; Martinez-Vaquero et al. 2015), where interactions occur in three (decision-making) stages: (i) Before the interaction, agents choose whether to propose an agreement; (ii) When receiving a proposal, agents decide whether to accept or reject it; (iii) During the course of the (repeated) interaction, agents choose whether to play SAFE or UNSAFE, depending on whether the agreement was formed; and in the case of multiplayer games, also on how many players committed to following the safety precaution, because a minimum number of committing players may be required for an agreement to be formed and put into effect.

The regimentation assumption entails that agents who accept an agreement proposal yet dishonour it by defecting during the interaction (i.e. fake committers (Han et al. 2013)), always honour and pay a compensation. Removing this assumption, another stage or decision point will be added: (iv) After each round of the (repeated) interaction, agents decide whether to use any type of peer incentives, depending on the decisions made by co-players and by themselves in the previous three stages.

This new decision-making stage adds extra layers of complexity, not just on how commitment behaviours are influenced by incentives, but also on how different types of incentives interact. Namely, we need to distinguish between incentives when an agreement is in place (i.e. incentives for agreement fulfillers or for violators), and when agreement is absent (i.e. incentives for mere cooperators or defectors). Considering all strategic behaviours concerning how to use incentives in co-presence in a population, we will examine how these two incentive sorts should be treated differently to foster safety-agreement compliance. Moreover, in order to have a clear understanding of the strength and weakness of each type of incentives, we will start from minimal models where only one type of (peer) incentive is present at a time. Increasingly more complex models, which include other types of incentive, will then be constructed and analysed to see how they interact and influence together the outcome of the race. So doing will provide deeper systematic understanding of how different types of incentives, whether separately or jointly, can be used to achieve high levels of safety compliance.

Rogue AI actors and teams might likely exercise antisocial incentives or bullying (AI-Roadmap-Institute 2017a; Herrmann, Thöni, and Gächter 2008). We will consider this possibility in our models. Antisocial or bullying teams most likely refuse to join a safety agreement and are detrimental to cooperation. However, setting up a pre-agreement can provide an efficient solution. One can implement measures to restrict access of non-participating (rogue) teams to AI knowledge (Han, Pereira, and Lenaerts 2016); while misbehaving participating teams can be appropriately handled through the agreement's terms and conditions, both by coordinated and institutional punishments (Sigmund et al. 2010; Boyd, Gintis, and Bowles 2010). This relates to the AI race, when a small number of major monopolizing actors or teams need to be confronted by the collective as a whole. Furthermore, promoting guilt (Pereira et al. 2017) may be envisaged as a further way to simmer down the AI race, viz. the recent developments regarding Facebook policy change promises and Cambridge Analytica's declared bankruptcy, by appealing to public discomfort and the chastising of data bullies, again, possibly with the help of crowd sourcing or inside employees.

Our analysis will first examine two-player games since they will permit us to focus on the effects of different types of peer incentives. We will then extend the analysis to multiplayer games, which will include strategic behaviours conditional on the level of participation of teams in a safety-agreement; that is, how many teams agree to comply with the safety requirements. Previous works have shown that increasing the number of players in interaction significantly magnifies the complexity of the evolutionary dynamics and outcome. Firstly, since the number of behavioural equilibria might significantly increase, it is important to study under what conditions and how likely is it that desirable equilibria can be reached; and, moreover, how different types of incentives should be used, separately or jointly, to improve the chance of reaching such desirable outcomes. Secondly, it appears that collective decisions are more difficult to be made for larger groups (Gokhale and Traulsen 2010). We will study how increasing group size influences the probability of safety-agreement formation and, once formed, which are the players' compliant behaviours. Indeed, as shown in previous works for regimented commitments within group interactions, it is crucial to closely monitor the minimum level of participation when deciding whether a commitment should be effectively formed in order to achieve an opti-

mal cooperation outcome (Han, Pereira, and Lenaerts 2016). Such a minimum membership requirement can be found in the creation of treaties that address international environmental issues and is also important for AI regulation agreements, as full consensus is rarely reached (as are the cases for cyberspace agreements organized by the UN (Taddeo and Floridi 2018)). Removing the regimentation, we will examine how different types and arrangements of incentives influence the participation level and how this level should be monitored to ensure safety-agreement compliance. Last but not least, in multiplayer agreements a new possibility arises where a group of players might coordinate to provide incentives with more substantial effects (e.g. gang up on free-riders). We will start by applying the existing models of coordinated and pool incentives (Sigmund et al. 2010; Boyd, Gintis, and Bowles 2010), where a group of incentive providers can share the cost of providing incentives and can decide to actually provide the incentives only if there is sufficient interest in sharing the cost.

In both pair-wise and multiplayer games, we will closely examine how incentives should be used differently, taking into account the diverse risk levels or probabilities that an AI disaster will occur. Does a high risk require stronger (or less so) incentives to ensure a high level of safety compliance? Also, considering that teams might have different capacities (for AI development), the question arises of how incentives should be used differently in light of a given capacity. For instance, stronger punishment might be required against teams who frequently violate safety requirements, especially if they are close to the finish line of the race. In contradistinction, stronger support and reward may be enacted for highly compliant teams, to ensure they win the race with a safe and powerful AI product.

Institutional incentives for safety-compliance

This final part of the agenda aims to examine how institutional incentives can be efficiently used for enhancing safety-compliance behaviours in the AI race games, and how they interact with peer incentives. The following questions will be addressed:

- What is the influence of different institutional incentives, separately or jointly, on safety-agreement compliance?
- How institutional incentives interact with peer incentives and how these two types can be jointly used to provide an efficient hybrid incentive strategy?

We will study institutional incentives in two scenarios: i) in the absence and ii) in the presence, of peer incentives. The former represents a fully centralised approach to regulating the AI development race, while the latter is a hybrid of centralised and decentralised regulations. In this latter case we explore different ways in which the two types of incentives interact. For example, being autonomous entities, the teams can decide by themselves the type of incentives used to enforce the safety-agreement, i.e. a peer or institutional incentive. It is natural to ask which option would emerge as the preferred one in the population and when. This can be answered by analysing EGT models in which all types of incentives are allowed to be adopted by agents in the population. Since forming a regulating institution is usually costly (Hilbe et al. 2014), we will identify when peer incentives are sufficiently efficient so that the institutional setting-up and maintenance cost can be avoided. This outcome is particularly important given the lack of attention so far to peer incentives for behavioural regulation in normative MAS, and moreover, to avoid (institutional) overregulation of AI development from the start.

In addition, a distinctive feature of the institutional setting is that the institution can have access to some global information, such as the current population composition. We explore approaches exploiting this distinctiveness. We will start by extending and generalising the 'first carrot, then stick' policy. The challenge is that there will be multiple types of strategic behaviours to incentivise, compared to only two types as in the usual standard institutional incentive models. At each time step, the institution needs to decide whether to incentivise (subject to a given budget) one type or even a subset of distinct types of incentive, depending on their current frequencies in the population. This institutional decision-making process is a complex multi-agent resource allocation problem (Han et al. 2018), for which appropriate resource allocation optimisation methods (e.g. from AI literature) can potentially be utilised.

Preliminary Results

In Supplementary Information, we described preliminary results for a two-team model of the AI race. The race is represented by a repeated game, consisting of a number of AI development rounds, where in each round teams can choose either to play SAFE or UNSAFE. The former choice is not just costlier, but also takes longer (i.e. slower speed of development). The team that wins the race will claim a significant benefit, unless AI disaster occurs. Our analysis considers a population of teams who can either play SAFE or UNSAFE in all the development rounds, or they can choose to adopt a reciprocal strategy (namely, conditionally SAFE). The teams interact and can adapt their strategy through social learning (i.e. copying the strategy of those who are more successful than them).

In general the analysis points to the direction that when the benefit from winning the race is high, teams that always choose UNSAFE dominate the population for a large range of parameters' values. This result shows that, in the context of the AI race with repeated interactions, the strategic nature and the outcome are different from those of standard repeated games. In the AI race the rogue teams can move faster in the race by ignoring safety precautions, and reciprocal strategies such as tit-for-tat still lose because of being nice initially. This initial finding suggests that, to drive the race in the more beneficial directions it is important to enact measures that influence (prohibit or accelerate) the speed of AI development of teams, since reciprocal behaviours might not be sufficient to promote cooperative or safety behaviours in this context.

References

[AI-Roadmap-Institute 2017a] AI-Roadmap-Institute. 2017a. Avoiding the precipice: Race avoidance in the development of artificial general intelligence.

[AI-Roadmap-Institute 2017b] AI-Roadmap-Institute. 2017b. Report from the ai race avoidance workshop, tokyo.

[Armstrong, Bostrom, and Shulman 2016] Armstrong, S.; Bostrom, N.; and Shulman, C. 2016. Racing to the precipice: a model of artificial intelligence development. *AI & society* 31(2):201–206.

[Baliga and Sjöström 2004] Baliga, S., and Sjöström, T. 2004. Arms races and negotiations. *The Review of Economic Studies* 71(2):351–369.

[Barrett 2016] Barrett, S. 2016. Coordination vs. voluntarism and enforcement in sustaining international environmental cooperation. *Proceedings of the National Academy of Sciences* 113(51):14515–14522.

[Baum 2017] Baum, S. D. 2017. On the promotion of safe and socially beneficial artificial intelligence. *AI & SOCIETY* 32(4):543–551.

[Bostrom 2017] Bostrom, N. 2017. Strategic implications of openness in ai development. *Global Policy* 8(2):135–148.

[Bostrum 2014] Bostrum, N. 2014. Superintelligence: paths, dangers, strategies.

[Boyd, Gintis, and Bowles 2010] Boyd, R.; Gintis, H.; and Bowles, S. 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328(5978):617–620.

[Cave and Ó hÉigeartaigh 2018] Cave, S., and Ó hÉigeartaigh, S. 2018. An ai race for strategic advantage: Rhetoric and risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*.

[Chen et al. 2015] Chen, X.; Sasaki, T.; Brännström, Å.; and Dieckmann, U. 2015. First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. *Journal of The Royal Society Interface* 12(102):20140935.

[Cherry and McEvoy 2013] Cherry, T. L., and McEvoy, D. M. 2013. Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics* 54(1):63–77.

[Geist 2016] Geist, E. M. 2016. It?s already too late to stop the ai arms race?we must manage it instead. *Bulletin of the Atomic Scientists* 72(5):318–321.

[Gokhale and Traulsen 2010] Gokhale, C. S., and Traulsen, A. 2010. Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U.S.A.* 107(12):5500–5504.

[Han et al. 2013] Han, T. A.; Pereira, L. M.; Santos, F. C.; and Lenaerts, T. 2013. Good agreements make good friends. *Scientific reports* 3(2695).

[Han et al. 2018] Han, T. A.; Lynch, S.; Tran-Thanh, L.; and Santos, F. C. 2018. Fostering cooperation in structured populations through local and global interference strategies. In *IJCAI-ECAI'2018*, 289–295.

[Han, Pereira, and Lenaerts 2016] Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2016. Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems* 1–23.

[Herrmann, Thöni, and Gächter 2008] Herrmann, B.; Thöni, C.; and Gächter, S. 2008. Antisocial Punishment Across Societies. *Science* 319(5868):1362–1367.

[Hilbe et al. 2014] Hilbe, C.; Traulsen, A.; Röhl, T.; and Milinski, M. 2014. Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *PNAS* 111(2):752–756.

[Hofbauer and Sigmund 1998] Hofbauer, J., and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[Martinez-Vaquero et al. 2015] Martinez-Vaquero, L. A.; Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2015. Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific reports* 5(10639).

[McCullough 2008] McCullough, M. 2008. *Beyond revenge: The evolution of the forgiveness instinct*. John Wiley & Sons.

[Nesse 2001] Nesse, R. M. 2001. *Evolution and the capacity for commitment*. Foundation series on trust. Russell Sage.

[Pamlin and Armstrong 2015] Pamlin, D., and Armstrong, S. 2015. Global challenges: 12 risks that threaten human civilization. *Global Challenges Foundation, Stockholm*.

[Pereira et al. 2017] Pereira, L. M.; Lenaerts, T.; Martinez-Vaquero, L. A.; and Han, T. A. 2017. Social manifestation of guilt leads to stable cooperation in multi-agent systems. In *AAMAS*, 1422–1430.

[Santos and Pacheco 2011] Santos, F. C., and Pacheco, J. M. 2011. Risk of collective failure provides an escape from the tragedy of the commons. *PNAS* 108(26):10421–10425.

[Shulman and Armstrong 2009] Shulman, C., and Armstrong, S. 2009. Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July*, 2–4.

[Sigmund et al. 2010] Sigmund, K.; Silva, H. D.; Traulsen, A.; and Hauert, C. 2010. Social learning promotes institutions for governing the commons. *Nature* 466:7308.

[Sigmund, Hauert, and Nowak 2001] Sigmund, K.; Hauert, C.; and Nowak, M. 2001. Reward and punishment. *P Natl Acad Sci USA* 98(19):10757–10762.

[Sigmund 2010] Sigmund, K. 2010. *The Calculus of Selfishness*. Princeton University Press.

[Singh 2013] Singh, M. P. 2013. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(1):21.

[Taddeo and Floridi 2018] Taddeo, M., and Floridi, L. 2018. Regulate artificial intelligence to avert cyber arms race. *Nature* 556(7701):296–298.

[Vasconcelos, Santos, and Pacheco 2013] Vasconcelos, V. V.; Santos, F. C.; and Pacheco, J. M. 2013. A bottom-up institutional approach to cooperative governance of risky commons. *Nature Climate Change* 3(9):797.

Supplementary Information – Preliminary Results

A Preliminary Two-player Model

Let's assume that in order to achieve AGI, a number of development steps or rounds are required, where in each round the development teams (or players) have two strategic options: to follow the safety precaution (SAFE) or to ignore this safety precaution (UNSAFE). Since it takes more time and effort to comply with the precautionary requirements, playing SAFE is not only costlier, but also implies a slower development speed, compared to playing UNSAFE. We assume that to play SAFE players need to pay a cost $c > 0$, while playing UNSAFE does not require them to pay any cost (or some cost smaller than c, which can be normalised to 0). Also, the development speed when playing UNSAFE is $s > 1$ (steps towards the powerful AI) while the speed when playing SAFE is normalised to 1.

Teams can collect benefits from their intermediate AI products. Assuming a fixed benefit, b, from the AI market, teams will share this benefit proportionally to their development speed. Moreover, we assume that with some probability p_{fo} those playing UNSAFE might be found out about their unsafe development and their products won't be used, leading to 0 benefit. Thus we can write the payoff matrix as follows (with respect to the row player)

$$\Pi = \begin{matrix} \\ \textit{SAFE} \\ \textit{UNSAFE} \end{matrix} \begin{matrix} \begin{matrix} \textit{SAFE} & \quad\quad \textit{UNSAFE} \end{matrix} \\ \begin{pmatrix} -c + \frac{b}{2} & -c + (1-p_{fo})\frac{b}{s+1} + p_{fo}b \\ (1-p_{fo})\frac{sb}{s+1} & (1-p_{fo}^2)\frac{b}{2} \end{pmatrix} \end{matrix}.$$

For instance, when two SAFE players interact, each needs to pay the cost c and they share the benefit b. When a SAFE player interacts with an UNSAFE one the SAFE player pays a cost c and obtains the full benefit b in case the UNSAFE co-player is found out (with probability p_{fo}), and obtains a small part of the benefit $b/(s+1)$ otherwise (i.e. with probability $1 - p_{fo}$). When playing with a SAFE player, the UNSAFE does not have to pay any cost and obtains a larger share $bs/(s+1)$ when not found out. Finally, when an UNSAFE player interacts with another UNSAFE, it obtains the shared benefit $b/2$ when both are not found out and the full benefit b when it is not found out while the co-player is found out, and 0 otherwise. The payoff is thus: $(1-p_{fo})\left[(1-p_{fo})(b/2) + p_{fo}b\right] = (1-p_{fo}^2)\frac{b}{2}$.

When a team achieves the objective of being the first to have developed AGI after having moved W steps, they obtain a benefit/prize B (which is shared among those who reach the target at the same time). However, an AI disaster might happen with some probability, which is assumed to increase with the number of times the safety requirements have been omitted by the winning team. When AI disaster occurs, the winning team loses all its benefits [1]. For simplicity, we assume that when no safety precaution is followed at all, the risk probably is given by $p_r \in [0,1]$, which is linearly decreasing the fewer times the winning teams violates the safety precautions. For example, if the team always abides to the precaution, then this probability is 0, while the team that only follows it half of the time has a risk probability of $p_r/2$.

Consider now the following three strategies acting in repeated steps in the AI development process:

- AS (always complies with safety precaution)
- AU (never complies with safety precaution)
- CS (conditionally safe, plays SAFE in the first round and then adopts the move its co-player used in the previous round)

The payoff matrix defining averaged payoffs for the three strategies assuming the first player wins after having moved W steps (denoted $p = 1 - p_r$)

$$\begin{matrix} \\ \textit{AS} \\ \textit{AU} \\ \textit{CS} \end{matrix} \begin{matrix} \begin{matrix} \quad\quad AS & \quad\quad\quad AU & \quad\quad\quad CS \end{matrix} \\ \begin{pmatrix} \frac{B}{2W} + \Pi_{11} & \Pi_{12} & \frac{B}{2W} + \Pi_{11} \\ p\left(\frac{sB}{W} + \Pi_{21}\right) & p\left(\frac{sB}{2W} + \Pi_{22}\right) & \Pi_{AU,CS} \\ \frac{B}{2W} + \Pi_{11} & \frac{s}{W}\left(\Pi_{12} + (\frac{W}{s} - 1)\Pi_{22}\right) & \frac{B}{2W} + \Pi_{11} \end{pmatrix} \end{matrix},$$

where just for the purpose of presentation, we denote $\Pi_{AU,CS} = p\left[\frac{sB}{W} + \frac{s}{W}\left(\Pi_{21} + (\frac{W}{s} - 1)\Pi_{22}\right)\right]$.

We have made some initial analysis of the population dynamics in a population of the three strategies AS, AU and CS, following EGT methods for finite populations (cf. Methods below), see Figure 1.

In general the analysis shows that when the benefit from winning the race increases, AU wins for an increasing range of risk probabilities. This result shows that, in the context of iterative AI development steps, the outcome differs from that observed in other repeated social dilemmas like the prisoners dilemma. In the AI bidding war the rogue teams playing AU can move faster in the race by ignoring safety precautions, and reciprocal strategies such as CS lose because of being nice initially (which is not the case in standard repeated games). Thus, to drive the dynamics towards the more beneficial directions (ensuring higher frequency of safety-compliance) it is important to put new mechanisms into place to control the speed of AI development of teams, because reciprocal behaviour might not be sufficient to promote cooperative or safety behaviours in this context. Based on the described model, when an UNSAFE act is revealed, we can consider different approaches to sanction the wrongdoing teams, e.g. peer versus institutional punishments, whether in presence or in absence of pre-commitments.

Note that in our model it is assumed that every team can afford to choose either SAFE or UNSAFE, regardless of how much funding they have for development. However, it could be the case that larger teams with a larger budget could play SAFE with even a greater speed than smaller teams choosing UNSAFE. As mentioned in the our

[1] In the current models we assume that an AI disaster might occur only when a true AI or AGI has been achieved, i.e. after the W development steps have been completed. However, it might be the case that some smaller scaled disasters might occur before that milestone, especially when it is not clear whether and when AGI will or has been be achieved, and there might even be false beliefs regarding its presence. What is more, parties may release over simplistic AI but deliberatively advertise more than it can achieve, thereby leading to unforeseen usage disasters. We will analyse these scenarios in future works.

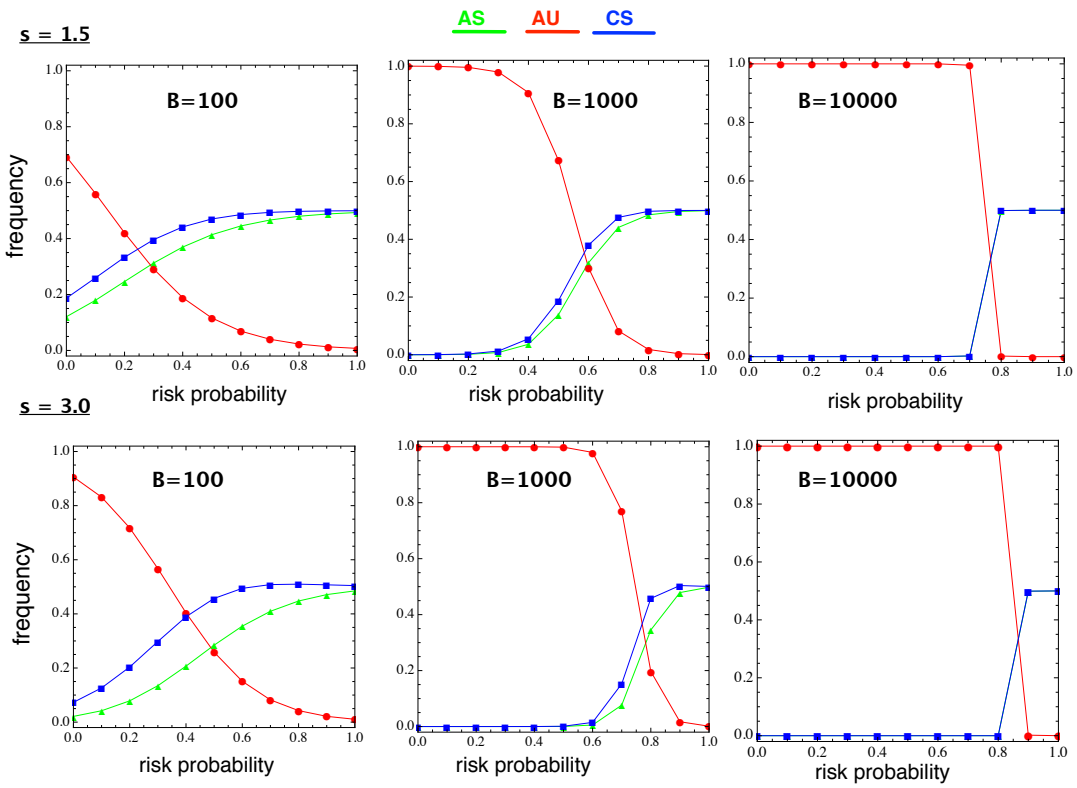

Figure 1: **Frequency of each strategy in a population of AS, AU and CS, as a function of the risk probability** p_r (i.e. that an AI disaster occurs when no safety precaution is considered). We plot for different alternative benefits in winning the AI race ($B = 100$ and 10000) and for different values of s (the number of steps corresponding to a UNSAFE move; $s = 1.5$ in the first row and $s = 3$ in the second row). In general, we observe that when the risk probability is small, AU is dominant. Also, the larger B and s, AU dominates for a larger range. Parameters: $c = 1$, $b = 10$, $W = 100$, $p_{fo}^2 = 0.1$, $\beta = 0.01$, population size, $Z = 100$.

Research Agenda, heterogeneity of teams' development capacities might significantly influence the dynamics and outcomes of the race. This issue will be incorporated into future models.

Methods: Evolutionary Dynamics in Finite Populations

Both the analytical and numerical results obtained here use EGT methods for finite populations (Sigmund 2010). In such a setting, players' payoff represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning, whereby the most successful players will tend to be imitated more often by the other players. Here social learning is modeled using the so-called pairwise comparison rule, assuming that a player A with fitness f_A adopts the strategy of another player B with fitness f_B with probability given by the Fermi function, $P_{A,B} = \left(1 + e^{-\beta(f_B - f_A)}\right)^{-1}$, where β conveniently describes the selection intensity. In Figure 1, the long-term frequency of each strategy in a population of several strategies in co-presence can be computed by calculating the stationary distribution of a Markov chain where its states represent the strategies in the population. Details of this calculation can be found e.g. in (Sigmund 2010; Han et al. 2013).